DOCUMENT RESUME

ED 438 501 CS 013 783

AUTHOR Heistad, David
TITLE Teachers Who Beat the Odds: Value-Added Reading Instruction in Minneapolis 2nd Grade Classrooms.
PUB DATE 1999-04-00
NOTE 43p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Canada, April 19-23, 1999). Appendices missing from original document.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Grade 2; Longitudinal Studies; Primary Education; *Reading Instruction; *Student Characteristics; *Teacher Effectiveness; *Teacher Influence; Teaching Styles
IDENTIFIERS Balanced Reading Instruction; Direct Instruction; Value Added Model

ABSTRACT

A study utilized data from four successive school years to explore the stability and correlates of teacher effects in reading. A "value-added" model was used to isolate the effect of the teacher from child demographic variables such as race, poverty, gender, family composition, and special learning needs. Teacher effects in second grade reading were found to have moderate stability over two consecutive years. Estimates of teacher effect stability increased substantially when value-added effects were aggregated over three or more years. Teacher effects in second grade reading correlated dependably with several facets of direct instruction philosophy and practice on a self-report survey. Teachers who demonstrated the highest value-added tended to disagree with the statement "reading and writing develop naturally, like speaking." They endorsed more use of small group instruction and more use of guided practice. Teachers identified as "exceptional" through value-added analysis endorsed more teacher directed activities, more development of word attack strategies and more use of individual student oral reading. Use of systematic motivational strategies and some form of test preparation activity were also endorsed to a greater extent by teachers with high value-added estimates. These findings are consistent with National Research Council findings on prevention of early reading difficulties. A balanced reading approach which utilizes explicit reading skill instruction was associated with higher reading success in second grade classrooms in this study. (Contains 61 references, 5 notes, and 20 tables of data.) (RS)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 438 501

TEACHERS WHO BEAT THE ODDS:

VALUE-ADDED READING INSTRUCTION

IN MINNEAPOLIS 2ND GRADE

CLASSROOMS

David Heistad, Ph.D.

Research, Evaluation and Assessment

Minneapolis Public Schools

Paper presented at the annual American Educational Research Conference

Montreal, Canada April 1999

CS 013 783

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D. Heistad

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ABSTRACT

Prior studies of teacher effects in reading have used models which fail to fully account for student background differences and prior reading ability. Studies of the stability of teacher effects conducted in the 1970s found low to moderate stability of student gain scores across successive academic years. Teacher behaviors which maximized academic engaged time were found to correlate dependably with teacher effects.

This study utilized data from four successive school years to explore the stability and correlates of teacher effects in reading. A "value-added" model was used to isolate the effect of the teacher from child demographic variables such as race, poverty, gender, family composition, and special learning needs. Teacher effects in second grade reading were found to have moderate stability over two consecutive years with median correlations varying from .4 to .6 depending on the number of students with pre and post test scores in a classroom. Estimates of teacher effect stability increased substantially when value-added effects were aggregated over three or more years.

Teacher effects in second grade reading correlated dependably with several facets of direct instruction philosophy and practice on a self-report survey. Teachers who demonstrated the highest value-added tended to disagree with the statement "reading and writing develop naturally, like speaking." They endorsed more use of small group instruction and more use of guided practice. Teachers identified as "exceptional" through value-added analysis endorsed more teacher directed activities, more development of word attack strategies and more use of individual student oral reading. Use of systematic motivational strategies and some form of test preparation activity were also endorsed to a greater extent by teachers with high value-added estimates.

These findings are consistent with National Research Council findings on prevention of early reading difficulties. A balanced reading approach which utilizes explicit reading skill instruction was associated with higher reading success in second grade classrooms in this study.

INTRODUCTION

Accountability for successful reading instruction is at the forefront of the American educational agenda. There is increasing pressure from state and federal governments, local civic groups, parent groups, and the general public to document the effectiveness or ineffectiveness of early literacy instruction. The public outcry over the gap between high academic standards and the present levels of reading has increased pressure upon American teachers to accelerate reading progress for young children (Snow, Burns & Griffin, 1998; Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Foorman, Fletcher, Francis, Shatschneider, and Mehta, 1998). Hence, a new era of teacher accountability has been initiated (Dwyer & Stufflebeam, 1996; Furhrman & O'Day, 1996; Berliner & Biddle, 1995; Kelly, 1997; Olson, 1998).

Educational indicators like the National Assessment of Educational Progress (NAEP) are not useful tools for holding teachers accountable for reading achievement (Meyer, 1994) because assessments are given too infrequently (i.e. every four years) and growth in reading cannot be localized to a group of students continuously enrolled within a particular classroom. The average test score on an assessment like the NAEP fails to account for mobility, student characteristics and the achievement level of students entering the classroom.

Since the average test score at a single point in time is inappropriate as an indicator of instructional effectiveness, some states and districts have used achievement gain on standardized tests for holding schools and teachers accountable. Some researchers have criticized the fairness and accuracy of simple gain indices (Berk, 1988; Glass, 1990). Statistical problems in the use of gain scores (i.e., correlation of initial status with gain) and uncontrolled family background, student ability, and past student achievement variables were sited as obstacles to the use of student performance on standardized tests to gauge the instructional effectiveness of teachers.

Districts and states where comprehensive statistical models were used to determine the unique contributions of schools or teachers to student performance (Dwyer &

Stufflebeam, 1996; Koretz, 1996) may have overcome these obstacles. Several researchers (Hanushek & Jorgenson, 1990; Willett, 1988; Meyer, 1994 & 1996) have indicated that the use of a value-added indicator that controls for prior achievement, student characteristics, and other non-classroom factors overcomes the problems with indicators based on average test scores or simple gain scores.

Use of value-added indicators for teacher accountability presupposes an accounting of the reliability of such indicators. One important test of teacher value-added indicators is the stability of such measures over time. As Brophy (1973) indicated, only after the stability of teacher effectiveness has been established and effects of within classroom cohorts controlled can the data from achievement tests be used for teacher accountability.

Once the stability of teacher effects is established, it is then useful to discern patterns of teacher behaviors that are associated with these effects. This type of investigation was the hallmark of the teacher effectiveness studies conducted by Brophy and his colleagues in the 1970s and early 1980's (Wittrock, 1986). Renewed investigation of teacher effectiveness indices and their stability could lead to another round of empirically based reading effectiveness studies that could shed light on teacher effects in the aftermath of the latest reading wars (Foorman, 1995; Lemann, 1997, Snow, et. al., 1998; Slavin & Fashola, 1998). According to Adams (1996, p. 16), the time has come for policy based on the scientific study of reading to replace the "theory-based educational reform" of the 1980's and 90's.

This study selected a form of the value-added indicator which was found to be accurate and unbiased (Meyer, 1996) to measure the effects of second grade reading instruction. In controlling for student characteristics such as poverty, gender, race, special learning program status and prior achievement, the indicator was used to distinguish instructional effects from external factors outside the influence of the teacher. The relationship between teacher value-added effects and reading instruction philosophy and practice related to whole language and skills-based approaches to reading instruction was also investigated.

REVIEW OF LITERATURE

Brophy, Evertson, and their colleagues completed a series of studies in the 1970s, beginning with an assessment of the stability of teachers' effects on achievement. Brophy (1973) obtained reading achievement scores for three consecutive years on the students of 88 experienced second grade teachers. Adjusted gain scores were calculated separately for word knowledge (vocabulary), word discrimination and reading comprehension. These adjusted gain scores (equivalent to a value-added indicator with only pre-test score as a predictor) were averaged across students for each teacher. Stability coefficients were low to moderate (most were in the .30s).

The earliest study of teacher effect stability found through literature search (Brown, 1971) reported a Spearman rank order stability coefficient of .55 for 54 first grade reading teachers in a metropolitan school district. This two-year study found no correlation between teacher effects and teacher experience, age, or education. Acland (1976) reported moderate stability coefficients for fifth grade teachers in word knowledge (.488), and language (.398), but somewhat lower coefficients for reading comprehension (.198), and language study skills (.132). These studies combined with the four studies reported by Rosenshine (1970) with stability coefficients ranging from -.08 to .53 encouraged Brophy and colleagues to embark on a series of studies of elementary teachers who were consistently high or low in student achievement effects (Brophy & Good, 1986).

Teacher effect stability was also investigated at the middle school level. Emmer, Evertson & Brophy (1979) studied 39 seventh and eighth grade English teachers over consecutive years and found considerable stability (an intra-class correlation of .55). Reading post-test scores adjusted for pre-test scores were analyzed for four consecutive years in the Texas Teacher Effectiveness Study (Brophy & Evertson, 1974). Analyses of trends over time indicated that about half of the 165 teachers in the study were stable in achievement effects. Thirty-one of these stable teachers were observed for 10 hours over the course of the first year of the study and 28 were observed for 30 hours during the second year. The results indicated that outstanding teachers managed their time efficiently,

assigned work at the appropriate difficulty level for individual students and used methods of positive reinforcement (Brophy & Good, 1986).

Correlates of Teacher Effects (1970s and 1980s)

The most consistent teacher effect correlating with adjusted achievement gain in the Brophy studies was academic engaged time (Rosenshine & Stevens, 1986). Teacher behavior correlated with high student engagement included a business-like orientation and high task orientation. Outstanding teachers tended to spend more time in guided practice, asked more questions and ensured a high percentage of correct responses.

A series of studies specifically focused on first grade reading (Evertson, & Brophy, 1979 & 1982) found that achievement gains were greater under the following conditions. More time was spent in reading groups and in active instruction, and less time was spent dealing with student misbehavior. Teachers managed classroom time efficiently by limiting transitions, sitting with small groups so as to be able to monitor the rest of the class, introducing lessons with overviews, and ordering student responses rather than allowing students to call out. Teachers who showed the greatest achievement gain presented lessons with frequent opportunities for students to read and answer questions about reading; they presented new words with explicit review of relevant phonics cues; and they made sure students work assignments were clear and would have students demonstrate how to do assignments before being released to work independently (Brophy & Good, 1986).

Value-added Studies

Meyer (1996) has articulated the rationale, theory and evidence for a system of value-added indices of school and teacher effects. "The key is to isolate statistically the contribution of schools from other sources of student achievement. This is particularly important in light of the fact that differences in student and family characteristics account for far more of the variance in student achievement than school-related factors." (p.200)

Meyer pointed out that average test scores for a single grade at a specific point in time reflect the learning that has taken place across a number of years. These test scores are

contaminated by the learning that took place prior to 1st grade or the time the student entered the class or school under consideration. The average test score is misleading because of four additional reasons:

1) Effects of student, family and community characteristics are confounded with instructional effects.
2) Average test scores reflect information about school performance which tends to be out of date. For example, student performance in eighth grade may be largely determined by instruction that takes place in early elementary school.
3) Average test scores tend to be contaminated by student mobility in and out of different schools. And mobility rates vary considerably from school to school.
4) Unlike the value-added indicator, the average test score fails to localize school performance to a common unit such as the classroom or grade level and thus is relatively weak as an accountability instrument (Meyer, 1996, p. 214).

A stability study of value-added effects for a large state-wide data base was conducted in South Carolina (Mandeville & Anderson, 1987; Mandeville & Rivers, 1991). Total reading and total math scores were obtained for all students in grades one through four on the Comprehensive Test of Basic Skills (CTBS). Student level data was aggregated to the grade level within each school. Post-test average scores were regressed on pre-test averages and the percentage of students eligible for free or reduced price lunch for each grade within the school. Within grade stability coefficients ranged from .34 to .66, depending on the grade level (Mandeville, 1988). However, Mandeville found that school effectiveness indices reflecting the performance of students at different grade levels were very unstable. In conclusion, he suggested that "grade-within-school effects dominate whatever global school effects operate in elementary schools" (Mandeville, 1988, p. 349). He did not, however, speculate whether teacher effects within grade level might also be found to be more stable than grade level effects.

The state of Tennessee developed an accountability system for schools and for teachers over ten years (Sanders & Horn, 1994) known as the Tennessee value-added

assessment system (TVAAS). TVASS analyzed the scale scores on the norm-referenced items in the Tennessee Comprehensive Assessment Program (TCAP). At the time of publication, the TVASS data base contained more than 3 million student records. The chief purpose of this system was to provide yearly reporting on school effects using a linear growth model. Scale scores on the TCAP were used to model a learning profile for each student. These profiles were grouped by district or school and produced a linear growth estimate for the school or district. The slope of these gains was compared to national norms and state expectations. Schools and school systems could then "identify where students are achieving normally, outstandingly, and substantially" (Sanders & Horn, 1994).

A recent longitudinal analysis of teacher effects by Sanders found that groups of students with comparable achievement scores in grade two had markedly different scores by grade five, and the difference was attributed to the quality of their teachers. Sanders indicated, "the single greatest effect on student achievement is not race, it's not poverty, it's the effectiveness of the individual classroom teacher" (Olson, 1998, p. 31).

The Tennessee researchers worked with the Tennessee Department of Education with the overall goal of holding individual teachers accountable for students' achievement. The team working on the Dallas system (described below) disagreed with this approach, indicating that it might lead to counter-productive competition between teachers in the same school (Dwyer & Stufflebeam, 1996).

Dallas Public Schools developed a somewhat different value-added indicator to measure classroom effects fairly and to hold schools, principals, and teachers accountable for student growth (Webster & Olson, 1988; Bembry, Webster et. al., 1994; Weerasinghe, & Mendro, 1997). Schools in Dallas used the classroom effectiveness indicators to analyze the effectiveness of individual classroom teachers. They concluded:

> "It is clear that teachers have large effects on student achievement, that effects have strong additive components over time, and that teacher effects are large enough to dwarf effects associated with most other interventions." (Bembry, Jordan, Gomez, Anderson, & Mendro, 1998)

The Early Reading Debate and Teacher Effects

Proponents of a whole language approach support the position that reading, like speaking, develops in a natural way and that classroom reading instruction should be child-centered, allowing students to construct their own personal knowledge of literacy through exploration (Goodman, 1970; Smith, 1971 & 1979). Goodman and Smith portray the good reader as skilled in the use of contextual information apart from simply processing letters. Teachers implementing the whole language approach tend to include shared reading activities to draw student's attention to word forms, letters, sounds, making predictions, and finding key ideas in the text (Foorman, et al., 1998). There is an emphasis on early writing with invented spelling, language extension activities, and integration of speaking, listening, reading and writing around themes (Eldredge, 1991). Whole language proponents tend to eschew skill sequencing, direct instruction of phonics, and teacher directed instruction (Stahl & Miller, 1989).

Critics of the whole language approach indicate that there are effective forms of direct instruction which are either ignored or actively opposed by whole language proponents (Pressley, 1994). Gough and Hillinger (1980) wrote an article entitled, "Learning to Read: An Unnatural Act" to counter the whole language contention that learning to read is as natural as learning to speak. Foorman (1995) indicated that humans are biologically specialized to produce oral language, but not so with reading and writing. Stanovich (1986) pointed out that it is not the good reader, but the least skilled reader who uses context in lieu of decoding. Chall (1983) followed up on her comprehensive study of the "great debate" in reading by offering evidence for stages of reading development. The second reading stage (grades 1 and 2) is called "the decoding stage" (p.15). The prescription for instruction at this stage is explicit instruction in the alphabetic principal, decoding skills instruction and also extensive oral reading. Adams' (1990) synthesis of basic reading research and field-level classroom research found that explicit code oriented instruction is critical for many students but so is extensive reading practice and exposure to a lot of reading materials.

A recent study of early reading teaching methods (Foorman, et al., 1998) contrasted teachers trained in direct instruction ("explicit phonics approach"), whole language classrooms ("implicit phonics approach") and a third group of teachers trained in an approach called "embedded phonics" (Hiebert, Colt, Catto, and Gary, 1992). Changes in vocabulary, phonological processing, and word-reading skills were assessed four times during the year for 285 first- and second-grade students. Results were analyzed using a three level HLM method with time nested within student and student nested within teacher. Teacher effects controlled for age, ethnicity and verbal IQ. The researchers found that students whose teachers instructed via the direct code approach improved reading and word recognition skills at a higher rate than students of whole language teachers (Foorman, et al., 1998).

Another recent study (Pressley et al., 1996) asked reading supervisors (randomly selected from the International Reading Association) to identify outstanding primary reading teachers who were effective in "educating large proportions of their students to be readers and writers" (p.366). This study of 123 teachers from across the country included detailed observations and surveys. The results of this research indicated that "exceptional teachers" reported: a) modeling of reading for students on a daily basis; b) practice and repetition of isolated skills with skill sheets, computers, songs, etc.; c) a combination of whole-group, small group, and individual instruction, including individual seat work; d) individual pacing of student work; e) integration of reading with the rest of the curriculum; f) continuous monitoring and student self-regulation. In discussing these findings, the investigators indicated that the outstanding teachers surveyed in this study engaged in both activities promoted by whole language and in code-oriented approaches - - thus, a balanced approach. Unfortunately, this study lacked a direct measure of reading growth to validate supervisor opinions regarding the effectiveness of reading instruction.

METHODS

This study was designed primarily to establish a measure of teacher effectiveness in reading and to investigate the degree of stability of that measure over time. A second purpose of the study was to investigate teacher philosophy, opinions and instructional behaviors associated with effective early reading instruction. Methodologically, this study was designed to isolate teacher effects from other sources of achievement variance so that instructional variables associated with reading achievement would be identified while controlling statistically for student characteristic differences.

Specific Hypotheses Tested

I. Teachers show no stability in 2nd grade reading instruction over successive years.

II. Teacher effectiveness is independent of self-reported teacher philosophy and general practices regarding reading instruction..

III. Teacher effectiveness is independent of reported use of direct instruction techniques and whole language approaches to reading instruction.

IV. Teacher effectiveness is independent of reported use of commercially developed test preparation materials.

V. Teacher effectiveness is independent of teacher length of service and academic credits earned.

Definition of Terms

Teacher effectiveness in this study was operationally defined as the individual teacher "value-added" (Meyer, 1996) regression coefficient η_s in the following general equation:

$$\mathbf{PostTest}_{is} = \gamma + \theta\,\mathbf{PreTest}_{is} + \alpha\,\mathbf{StudChar}_{is} + \eta_s + \varepsilon_{is}. \quad \text{Equation 3.1}$$

where i indexes individual students and s indexes teachers; PostTest_{is} and PreTest_{is} represent student reading achievement for a given student in second grade and first grade, respectively; StudChar represents a set of individual and family characteristics assumed to determine growth in student achievement growth; ε_{is} , the error term, captures the unobserved student-level determinants of achievement growth; γ is a constant; θ and α are

model parameters that must be estimated; and η_s is the teacher effect that must be estimated. Teacher effects, calculated through this equation, represent the contribution of a given teacher to growth in student achievement after controlling for all student-level factors.

Student characteristics in this regression equation are defined as follows:

α_1 = Free or reduced price lunch - coded "1" for free or reduced price lunch; "0" for full price lunch;

α_2 = Resides with - coded "1" for lives with two parents; "0" for other living arrangements including single mother, single father, relative, by self;

α_3 = Limited English Proficient (LEP) - coded "1" enrolled at the time of post test in Limited English Proficiency Programs; "0" non LEP;

α_4 = Special Education - coded "1" for current individual education plan (IEP) at the time of the post test; "0" for no current IEP;

α_5 = African American - coded "1" for enrolled as "African American" for; "0" enrolled as Asian, Hispanic, White or American Indian;

α_6 = American Indian - coded "1" for enrolled as African American; "0" enrolled as Asian, Hispanic, White or American Indian.

All student characteristic codes were downloaded from the Minneapolis School District mainframe computer. Year I was academic year 1993-94; Year II was academic year 1994-95; and Year III was academic year 1995-96. Descriptive statistics for the population and sample are presented in Table 1.

This study was conducted with approval from Minneapolis Public Schools (MPS) central office personnel and the president of the Minnesota Teacher's Federation, Local 59 which represents MPS teachers in collective bargaining. In accordance with this agreement, all teacher names were kept strictly confidential. Several sources of information were used to verify teacher assignments to homerooms during the three years of the study. Teacher rosters collected from every school were cross-referenced with the district staff directory of teachers assigned to each school. For Year 2 and Year 3, the homeroom field coded on the standardized testing data tape received from the test publisher was used as a third source to verify these data.

Table 1. Minneapolis Public School Second Grade Population and Second Grade Study Sample 1995-1996

Category	District Number	Percentage	Study Sample Number	Percentage
American Indian	254	6.1%	189	5.8%
Asian American	516	12.4%	384	11.9%
Hispanic	201	4.8%	95	2.9%
African American	1823	43.6%	1341	41.4%
White American	1383	33.1%	1228	37.9%
Free or reduced price lunch	2803	67.1 %	2086	64.4%
Resides with both parents	2058	49.2%	1663	51.4%
Limited English Proficient	505	12.1%	286	8.8%
Special Education	386	9.2%	303	9.4%
Total	4177	100 %	3237	100 %

The stability analysis study sample consisted of all teachers in the Minneapolis Public Schools who taught second grade for two consecutive years, 1993-94 and 1994-95 or 1994-95 and 1995-96. Teachers who changed schools during this period were included in this analysis as long as they continued to teach second grade. Stability analysis was conducted on teachers who had at least seven second grade students in their class for two consecutive years. Those classes having more than one teacher during the school year or where teacher assignment could not be verified were also excluded from the study. Table 2 indicates the number of teachers who met the inclusion criterion by classroom cohort size for each of the three study years.

Table 2. Number of 2nd grade teachers by classroom cohort size

Cohort Size	Less than 7	7-10	11-14	15-18	more than 18	Total Study Sample
Number of Teachers 1993-94	56	33	58	75	16	182
Number of Teachers 1994-95	49	47	70	69	11	197
Number of Teachers 1995-96	49	43	68	71	24	206

Assessment Instruments

Tests selected to measure reading comprehension achievement were the California Achievement Tests, Form E (CAT/E), reading comprehension and vocabulary subtests Levels 10, 11 and 12 and the California Achievement Tests, Fifth Addition (CAT/5), reading comprehension subtests Levels 10, 11, and 12.

A three-part teacher survey was constructed to assess reading instruction strategies, general philosophy of reading instruction, and use of test preparation activities for teachers who instructed second grade students during the 1996-97 school year. The first page of this survey was adapted from a reading study conducted by Doug Marston, a Minneapolis School Psychologist. The 26 items on the original survey were examined with a factor analysis and found to have two main factors: one with direct-instruction/phonics type items (i.e. initial guided practice, individual oral reading, explicit phonics instruction, frequent & direct progress monitoring, present material in small steps, development of word attack strategies, develop sight vocabulary); and the other was a whole language & reading/writing process factor (i.e. shared book experiences, journal writing, emphasize meaning during reading, encourage prediction during reading, literature extension activities, share published books/projects, collaborative writing). Four items from the original survey were eliminated because they correlated equally with the two main factors. The final survey was formatted for scanning with an electronic scanning machine.

The second page of the survey dealt with general reading instruction practices and philosophy. These items were filled out with the whole class in mind and the teacher was asked to mark each response on a line 100 centimeters long to questions related to instructional grouping practices, degree of teacher direction, and philosophy of reading instruction. Following this section, 3 questions regarding use of test preparation materials were asked. On the third page of the survey, each teacher was asked what, if any, published test preparation materials were used prior to the previous year spring achievement testing[1].

[1] See Appendix C for a copy of the survey sent to all second grade teachers

Procedures

The following data were gathered on all second grade students and their teachers for three consecutive years 1993-94, 1994-95, 1995-96.

Student data:
Sex
Free or reduced price lunch status
Zip code
Racial/ethnic category
Parent or guardian "resides with" status
Limited English Proficiency (LEP) status
Special Education status
California Achievement Test spring reading scores

Teacher data:
Homeroom
Years of teaching experience
Number of graduate education credits
Survey of specific reading strategies (22 items)
Survey of reading instruction philosophy (6 items)
Survey of test preparation practices (3 items)

California Achievement Test Reading Comprehension raw scores were converted to Normal Curve Equivalent units[2] by the test publisher and linked with individual student names by the unique district student identification number. Spring testing files for successive years were matched to form classroom cohorts. Demographic data for each file were taken from a mainframe download during January of each school year when special education status and free or reduced price lunch status were finalized for government reporting purposes.

Three sources of information were used to verify teacher assignments to homerooms during the three years of the study. Teacher rosters collected from every school were cross-referenced with the district staff directory of teachers assigned to each school. These data were verified using the homeroom field coded on the standardized testing data received from the test publisher and the homeroom field downloaded from the district's Unisys mainframe computer. Years of teacher experience and number of graduate education credits were downloaded from the district Human Resources Department data base and matched to teacher name and home room.

Surveys were distributed at the end of January 1997 to all 255 second grade teachers in the Minneapolis Public School District. The district Superintendent of Schools wrote a

[2] Normal Curve Equivalent (NCE) is a standard score with mean of 50 and standard deviation of 21.06 which is commonly used in evaluating Title 1 federal programs for disadvantaged students.

cover letter for the survey, encouraging full participation. Thank you notes and reminders were sent to teachers in order to maximize response rates. In total, 186 (73%) teachers returned completed surveys. All survey responses were merged with value-added teacher effects for the 1995-96 classroom cohorts. Of the 186 respondents, 80 teachers did not provide primary reading instruction to second grade students in 1995 & 1996 or had less than 7 students tested during both years. The remaining 106 teacher surveys were included in statistical tests of the key research questions relating to instructional practices and general reading instruction philosophy.

Statistical Analysis

Teacher effects were calculated separately for each of the three study years, 1993-94, 1994-95 and 1995-96 using the multiple regression procedures for value-added outlined by Meyer (1996). All student demographic factors were dummy coded '1' or '0'. A set of teacher dummy-variables were generated so that each teacher effect would appear as a coefficient in the regression analysis. All of the teacher dummy codes were entered simultaneously with the student characteristic dummy codes in a standard SPSS® (1993, version 6.0) regression analysis.

Tests of the hypothesis of no teacher effect stability were performed using Pearson product-moment correlations among the three study year's value-added coefficients. Teacher effect stability was further investigated with generalizability studies (G-studies) which included teachers with value-added coefficients for at least seven students over the three study years. With teachers as the facet of differentiation and occasions as the random facet, the generalizability coefficient calculated is equivalent to Cronbach's Alpha (Cronbach et al., 1972). Variance components estimated from these studies were used to project the increase in teacher effect stability over multiple occasions using a decision study (D-study). The relationship between value-added stability and the size of classroom cohort size was calculated and plotted for the two-year and three-year stability estimates.

Teacher effects from Year 3 were correlated with reading instruction survey results using Pearson Product-moment statistics for questions measured on an equal interval scale

(page 2, questions 1-6) and Spearmen rank-order correlations for questions measured on an ordinal scale (page 1, questions 1-22). All null hypothesis tests were performed at the conventional type 1 error rate of .05. Analyses were first conducted on all 106 teachers with value-added coefficients for Year 3. A second set of analyses was conducted on the 68 teachers with value-added coefficients for all three study years. Teachers who were consistently high in value-added were compared with other teachers using t-tests for interval level survey data and a non-paramentric tests of independent groups called the Mann-Whitney U test for rank order survey items.

RESULTS

The major purpose of this study was to determine the stability of teacher effectiveness in second grade reading instruction. A value-added regression coefficient was calculated for each teacher for each of three consecutive years. In Years 1 and Year 2, the CAT/E total reading normal curve equivalent (nce) score served as both pre-test and post-test reading indices. In Year 3, the CAT/5 reading comprehension nce was the post-test score and the CAT/E total reading nce was the pre-test score. Unstandardized regression coefficients for the pre-test and demographic variables are presented in Table 3. These coefficients are in nce units which have a mean of 50 and standard deviation of 21.06 in the standard normal distribution.

Table 3.Unstandardized regression coefficients for all three study years

Variable	1993-94	standard error	1994-95	standard error	1995-96	standard error
Constant	16.14	1.15	16.88	1.15	15.24	1.30
Total Reading pre-test	0.73	0.01	0.74	0.01	0.72	0.01
African American	-4.37	0.66	-4.79	0.65	-3.01	0.72
American Indian	-3.73	1.10	-5.19	1.09	-3.26	1.28
Gender	-1.92	0.50	-0.96	0.48	-2.86	0.52
Lives with 2 parents	0.80	0.61	0.12	0.58	0.55	0.63
Free/educed price lunch	-4.07	0.67	-5.00	0.67	-3.92	0.72
Resides in high poverty zip	-0.67	0.62	-0.78	0.65	-1.50	0.71
Limited English Proficiency	-6.59	1.17	-4.95	0.98	1.08	1.08
Special Education	-3.69	0.89	-5.09	0.78	-5.97	0.94

Value-added teacher effects were calculated using a dummy code "1" if the student was in the teachers classroom and instructed in reading during the specific year in question;

or "0" for each student who was not in this classroom. This method required considerably more computer resources but had the advantage of yielding individual standard errors for each teacher. For example, the 1994-95 regression analysis for each year included a single dependent variable, 9 student demographic independent variables, and 219 teacher dummy variables. For each classroom included in the value-added analysis, 218 of the teacher variables were coded "0" and one of the teacher variables (for the homeroom teacher the student was enrolled) was coded "1." Students in classrooms with less than 3 second grade students or in classrooms where the homeroom teacher did not provide reading instruction were coded "0" for all 218 teacher variables and thus provided a "virtual classroom" for comparison. Regression output included separate teacher effect standard errors for each teacher included in the analysis. The mean standard errors of the classroom effects decreased from 6.3 nces for four students in a classroom to 3.2 nces for 21 students in a classroom. Value-added effects for all 101 teachers included in the three-year analysis are presented in Appendix A.

Regression Prediction Validation

The stability of the demographic variable regression coefficients is evident from visual inspection of Table 3. Racial/ethnic coefficients for African American & American Indian ranged from about –3 to –5 nces. The free or reduced price lunch coefficient ranged from about -4 to -5 nces. Lives with both parents ranged from about +0 to +1 nce and lives in high poverty zip code ranged from about -.5 to –1.5. The coefficient for Special Education decreased from year 1 (-3.7 nce) to year 3 (-6.0 nces) while the coefficient for LEP increased from about –6.5 to +1.0 over the same period.

Statistical analyses were performed to further establish the consistency & predictive power of the regression equation. Hierarchical multiple regression was used to determine the degree to which student demographic characteristics contributed to the prediction of the reading post-test score over and above the pre-test score. Table 4 indicates the increase in R^2 with the addition of racial/ethnic variables, gender, family composition and poverty, and special program status. A cross-validation of the full regression formula was performed on

Year 1 data using the Year 2 coefficients and conversely on the Year 2 data using the Year 1 coefficients. Very minimal shrinkage in R^2 was found in this double cross-validation. In Year 1 the R^2 decreased from .662 to .659; in Year 2 the R^2 decreased from .694 to .690.

Table 4. Change in multiple regression R^2 with hierarchical inclusion of student variables

Variables	R^2 1993-94	R^2 1994-95	R^2 1995-96
Total Reading pre-test score	.632	.656	.560
Pre-test + race	.643	.667	.564
Pre-test + race + gender	.645	.668	.568
Pre-test + race + gender + family composition and poverty	.654	.686	.579
Pre-test + race + gender + family composition and poverty + special program status (full model)	.662	.694	.587
Cross-validation (full model)	.659	.690	-
Full model + teacher effects	.705	.750	.682

A step-wise inclusion procedure was used to determine which variables failed to add significantly to the prediction equation for each of the three study years. In Year 1 "resides with both parents" and "resides in high poverty zip code" failed to enter the step-wise regression. In Year 2 "resides with both parents" and gender failed to enter. In Year 3 only LEP status failed to enter the step-wise regression. Since no variable was consistently excluded using step-wise criteria it was decided to use the full model to determine teacher value-added effects. The magnitude of teacher effects is depicted in the last line of Table 4. Teacher effects added 4.3% to 9.2% post-test variance accounted for over and above the pretest and demographic variables in the model.

The Stability of Teacher Effects

Pearson product moment correlation coefficients were calculated for all three combinations; Year 1: Year 2, Year 2: Year 3 and Year 1: Year 3. Stability coefficients increased with the size of classroom cohorts, as noted in Figure 1 and Table 5. The median stability coefficient for 132 classrooms with at least 7 students in the pre-post classroom cohort for both years was .449 [$t_{(131)} = 5.46$; $p < .001$]. The median stability coefficient for 87 classrooms with at least 12 students in the cohort for both years was .519 [$t_{(86)} = 5.62$; p

$< .001$]. Even with much reduced sample size (n = 24), the median stability coefficient for classrooms with at least 16 students in the cohort was .604 [$t_{(23)} = 2.86$; $p < .021$].

These analyses permit rejection of the hypothesis that second grade reading effects are not stable across consecutive years. By rejecting the hypothesis, the dependability of value-added indicators of teacher effectiveness is supported.

Figure 1. Median Stability of Value-added Coefficient as a Function of Cohort Size

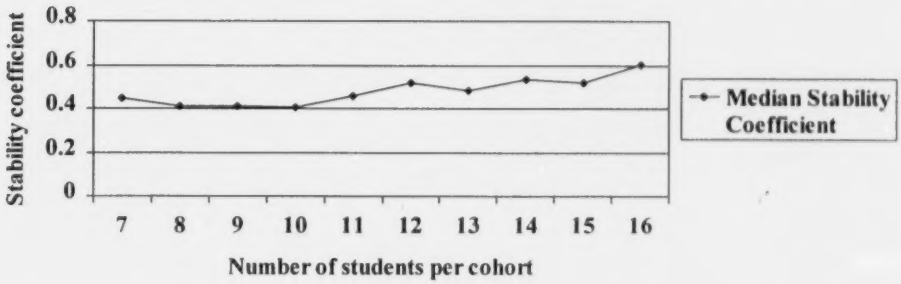


Table 5. Teacher effects stability as a function of the number of students with pre-test and post-test scores

Cohort size (no. of students)	Years 1 & 2 Stability[3]	Years 2 & 3 Stability Coefficient	Years 1 & 3 Stability Coefficient	Median Stability Coefficient
7 or more	.449	.381	.549	.449
	n= 120	*n= 132*	*n=116*	
8 or more	.412	.392	.560	.412
	n= 113	*n= 118*	*n= 113*	
9 or more	.411	.407	.557	.411
	n= 108	*n= 114*	*n= 111*	
10 or more	.406	.401	.554	.406
	n= 104	*n= 107*	*n= 103*	
11 or more	.458	.415	.516	.458
	n= 97	*n= 101*	*n= 101*	
12 or more	.519	.367	.526	.519
	n= 88	*n= 87*	*n= 93*	
13 or more	.518	.365	.482	.482
	n=77	*n= 76*	*n= 79*	
14 or more	.548	.392	.536	.536
	n= 63	*n= 59*	*n= 64*	
15 or more	.520	.465	.543	.520
	n= 44	*n= 50*	*n= 46*	
16 or more	.725	.526	.604	.604
	n= 24	*n= 26*	*n= 23*	

[3] n = number of teachers

BEST COPY AVAILABLE

A generalizability study (G-study) was conducted on the 101 teacher effects for classrooms with at least 7 students in each of the three study years. With teacher as the facet of differentiation and occasion as the random facet, the generalizability coefficient (similar to Cronbach's Alpha) was .737. In Table 6, the variance components for the teacher facet and teacher by occasion facet may be observed.

Table 6. Value-added teacher effects stability for 101 teachers included in the study for three consecutive years

Analysis of Variance			
Source of Variation	DF	Mean Square	Variance Component
Between Teachers	100	82.5869	20.28
Within Teachers	202	24.6945	
Occasions	2	.0000	0.00
Occasions x Teachers	200	21.7463	21.75
Total	302	43.8642	
Reliability Coefficients Alpha = .7367	3 occasions		
D-Study			
Generalizability Coefficient for 1 occasion ρ^2= .483			
Generalizability Coefficient for 2 occasion ρ^2= .651			
Generalizability Coefficient for 3 occasion ρ^2= .737			
Generalizability Coefficient for 4 occasion ρ^2= .789			

The generalizability coefficient, which is denoted ρ^2, is computed as the ratio of universe score variance to expected observed score variance (Brennan, 1983). In this G-study, teachers in the Minneapolis Public Schools who teach 2nd grade reading for three years, constituted the universe of generalization. Increase in the dependability of teacher effects were estimated using the G-study variance components in a D-study where changes in generalizability were computed as a function of increased number of occasions (Brennan, 1983, p.12). The D-study generalizability estimates in this study increased from .483 for a single 2nd grade cohort to .789 for four cohorts of 2nd grade students.

BEST COPY AVAILABLE

Teacher Effect correlates - Dimensional Analysis

Instructional behaviors, teacher opinion, and philosophy of reading instruction were investigated with teacher self-report surveys. Six items dealing with general instructional practices and philosophy of reading instruction were formatted with a 100-centimeter line and polar opposite descriptors (e.g. small group instruction 100% of the time vs. whole class instruction 100% of the time). Teachers were asked to mark with an "X" on the line indicating their position on the continuum.

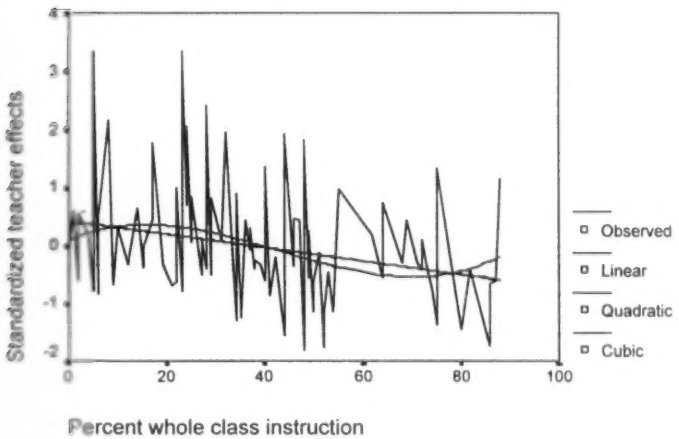
Trend analysis of teacher reading philosophy detected a significant correlation between teacher value-added coefficients and three of the six questions. Negative correlations were found between teacher effects and whole class grouping (see Table 7 and Figure 2).

Table 7. Percent whole class vs. small group instruction ANOVA Table

Contrast	R^2	d.f.	F	Sign.
Linear	.069	104	7.76	.006
Quadratic	.069	103	3.84	.025
Cubic	.090	102	3.37	.022

Figure 2.

Trend analysis on whole class instruction



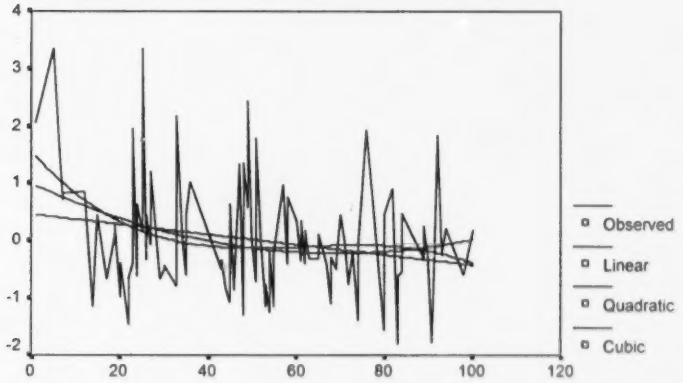
BEST COPY AVAILABLE

Negative correlations were also found between teacher effects and endorsement of the statement, "Reading and writing develop naturally, like speaking" (see Table 8 & Figure 3).

Table 8. Extent of agreement with the statement, "Reading and writing develop naturally, like speaking."

Contrast	R^2	d.f.	F	Sign.
Linear	.049	103	5.30	.023
Quadratic	.075	102	4.14	.019
Cubic	.091	101	3.39	.021

Figure 3. Trend analysis on the statement, "Reading and writing develop naturally"



0= strongly disagree; 100 = strongly agree

Twenty-two items dealing with specific reading strategies were rated on a four-point dimension from "none" to a "significant amount." This scale was assumed not to be equal interval, therefore results on this portion of the survey were analyzed with non-parametric methods. Spearman correlations between teacher effects and each of the 22 items are presented in Table 11.

BEST COPY AVAILABLE

Table 11. Correlation of specific reading strategy items with teacher value-added coefficients.

Variable	Median (1-4 scale)	Spearman Correlation	P value
Begin a lesson with a short review of previous learning	3.3	.025	.77
Shared book experiences	3.2	-.101	.30
Have student visualize while reading	2.6	-.035	.72
Independent reading	3.3	.108	.27
Modeling of reading for student	3.6	-.016	.87
Development of word attack strategies	3.5	.152	.12
Present new material in small steps, with student practice after each step	3.4	-.161	.10
Student reads non-fiction material.	2.9	.011	.91
Student shares his/her own published books/projects	3.1	-.045	.65
Individual student oral reading	3.4	.118	.22
Choral reading	3.4	.103	.29
Journal writing	3.2	.044	.66
Emphasize meaning during reading instruction	3.4	.062	.53
Guide student during initial practice	3.4	.192*	.05
Encourage prediction while reading	3.2	.030	.76
Develop sight vocabulary	3.5	.163	.10
Spelling homework and frequent spelling assessment	3.4	.023	.81
Whole language approach	3.0	-.263*	.01
Collaborative writing	2.6	.000	.99
Explicit and direct phonics instruction	3.5	-.022	.83
Monitor student reading progress directly and frequently	3.5	.125	.21
Literature extension activities	2.9	-.099	.32

* Effects are statistically significant at .05 type 1 error rate

Two strategies were dependably correlated with teacher effects, "Guide student during initial practice" (r=.192; p=.05) and "Whole language approach" (r=-.263; p=.01). Three other items approached statistical significance: "Development of word attack strategies" (r=.152; p=.12); "Develop sight vocabulary" (r=.163; p=.10) and "Present material in small steps, with student practice after each step" (r=-.161; p=.10). A separate item, "Have you used systematic motivational strategies to encourage improved reading achievement with this student?" also approached significance (r= .170; p= .08).

BEST COPY AVAILABLE

The Pearson product-moment correlation between value-added teacher effects and teacher years of service was non-significant (r=-.081; p=.48). Similarly, the correlation between teacher credits earned and value-added was non-significant (r=-.023; p= .84).

There was a dependable relationship between test preparation and teacher effects (r=.233; p=.02). However, there was no significant linear relationship between time spent in test preparation activities and teacher effects (r=.060; p=.56). Quadratic and cubic trend analyses were also performed and found to be non-significant (r=.02; p=.79 and r=.05; p=.74 respectively).

Categorical Analysis

Teacher effects for the 101 classrooms with at least 7 students in each of the three study years were used to categorize teachers in the top 20%. Teachers who appeared in the top 20% all three study years (6), and teachers who were in the top 20% two of three years (12), were termed "exceptional." These coefficients were matched with the file of returned surveys to form a file of 68 teachers: 11 "exceptional" teachers, and 57 "other" teachers.

The six items dealing with overall philosophy of reading instruction were analyzed with student t-tests. The means were found to be dependably different between "exceptional teachers" and "other teachers" for three of the six items.

Table 12. Differences between "exceptional" and "other" teacher on whole class instruction item.

Variable	Number of Cases	Mean	SD	SE of Mean
I1 Small group vs. whole class instruction				
Other teachers	55	40.2909	21.813	2.941
Exceptional teachers	11	26.0909	20.926	6.309

Mean Difference = 14.2000

Table 12 indicates that "exceptional teachers" reported an average of approximately 25% of the time spent in whole class reading instruction, while "other teachers" reported approximately 40% of reading instruction with the whole class. The mean difference of

BEST COPY AVAILABLE

14.2% was dependably different from zero [$t_{(65)} = 1.98$; $p = .05$]. Exceptional teachers were somewhat more likely to report that reading lessons are teacher directed versus student choice [$t_{(65)} = 1.55$; $p = .12$], but the difference was not statistically dependable at the conventional .05 Type 1 error level.

Table 13. Differences between "exceptional" and "other" teachers on teacher directed versus student choice.

Variable	Number of Cases	Mean	SD	SE of Mean
I2 Teacher directed vs. student choice				
Other teachers	55	24.3273	19.848	2.676
Exceptional teachers	11	14.7273	11.680	3.522

Mean Difference = 9.6000

There were approximately 14 points of difference (on the 100 point scale) between "exceptional teachers" and other teacher on the item, "Reading and writing develop naturally, like speaking" as presented in Table 14. This difference approached statistical significance [$t_{(65)} = 1.90$; $p = .06$].

Table 14. Differences between "exceptional" and "other" teacher on the question, "Reading and writing develop naturally, like speaking."

Variable	Number of Cases	Mean	SD	SE of Mean
I3 Reading and writing develop naturally, like speaking				
Other teachers	55	51.2727	22.084	2.978
Exceptional teachers	11	37.6364	19.765	5.959

Mean Difference = 13.6364

Responses to the reading worksheet item were relatively similar between exceptional and "other teachers". Both groups located on the "agree" side of the midline in response to the question, "There is nothing wrong with well-devised worksheets emphasizing letter-sound relationships and word analysis skills." Given the relatively large within group

BEST COPY AVAILABLE

variance on these items, the between group difference of 8.4 units was not statistically different [$t_{(65)} = 1.18$; $p = .24$].

Table 15. Differences between "exceptional" and "other" teacher on the reading worksheet question.

Variable	Number of Cases	Mean	SD	SE of Mean
I4 There is nothing wrong with well devised worksheets				
Other teachers	55	66.2364	21.073	2.841
Exceptional teachers	11	74.6364	24.373	7.349

Mean Difference = -8.4000

There was very large within group variance for "exceptional teachers" on the question referring to controlled vocabulary vs. authentic texts. Both groups tended to disagree with Goodman's (1989) statement, yet "exceptional teachers" tended, on average, to disagree less with the statement, "Meaningful, predictable authentic texts are incompatible with controlled vocabulary and decontextualized phonics instruction." The difference between exceptional and "other teachers" was not dependably different from zero in a separate variance t-test [$t_{(11.86)} = .98$; $p = .35$].

Table 16. Differences between exceptional and "other teachers" on the compatibility of controlled vocabulary with authentic texts.

Variable	Number of Cases	Mean	SD	SE of Mean
I5 Meaningful, predictable texts are incompatible with controlled vocabulary				
Other teachers	53	33.1509	23.090	3.172
Exceptional teachers	11	44.0000	35.086	10.579

Mean Difference = -10.8491

Both groups of teachers tended to agree with Chall's (1990) statement, "In second grade most students are at the stage of reading development where direct instruction in letter-sound relations (phonics) and practice in their usage is critical." Again the differences were not statically dependable.

BEST COPY AVAILABLE

Table 17. Differences between exceptional and "other teachers" on the necessity of direct phonics instruction.

Variable	Number of Cases	Mean	SD	SE of Mean
I6 Direct instruction in letter-sound relations is critical				
Other teachers	54	67.7963	22.237	3.026
Exceptional teachers	11	74.2727	36.233	10.925

Mean Difference = -6.4764

The 22 items dealing with specific reading strategies for randomly selected below-average students were analyzed using the non-parametric equivalent of the t-test, the Mann-Whitney U statistic & Wilcoxon Rank Sum statistics. Results of these comparisons, presented in Table 18 include dependable differences for development of word attack strategies, use of individual student oral reading, and explicit and direct phonics instruction.

Table 18. Differences Between Exceptional and Other Teachers on 22 Specific Reading Strategies for Low-Achieving Students.

Variable	Exceptional Teacher Mean Rank	Other Teacher Mean Rank	P value[4]
Begin a lesson with a short review	[illegible]	34.9	.69
Shared book experiences	30.3	35.3	.41
Have student visualize while reading	36.0	33.7	.71
Independent reading	40.0	33.4	.28
Modeling of reading for student	30.3	35.3	.38
Development of word attack strategies	**45.5***	**32.4**	**.03**
Present new material in small steps	33.9	34.6	.90
Student reads non-fiction material.	41.0	33.3	.20
Student shares own published books	31.6	35.1	.59
Individual student oral reading	**43.9**	**32.7**	**.06**
Choral reading	33.8	34.6	.29
Journal writing	40.2	33.4	.27
Emphasize meaning during reading	37.5	33.9	.55

[4] Results of Mann-Whitney sum of ranks statistics and approximate t-test

BEST COPY AVAILABLE

Guide student during initial practice	**44.6***	**32.6**	**.04**
Encourage prediction while reading	36.1	34.2	.75
Develop sight vocabulary	40.4	33.4	.25
Spelling homework and spelling assessment	39.5	33.5	.32
Whole language approach	35.9	34.2	.78
Collaborative writing	40.4	33.4	.25
Explicit and direct phonics instruction	**44.5***	**32.6**	**.05**
Monitor student reading progress directly	38.2	33.8	.46
Literature extension activities	30.9	35.2	.49

Teachers identified as exceptional reported using systematic motivational strategies 82% of the time while 51% of "other teachers" reported using systematic motivational strategies (see Table 19). This difference was statistically dependable [$t_{(66)} = 2.22$; $p = .04$].

Table 19. Difference between "exceptional teachers" and "other teachers" on use of systematic motivational strategies (coded 1 = yes, 0 = no).

Variable	Number of Cases	Mean	SD	SE of Mean
MOTIVATION				
Other teachers	57	.5088	.504	.067
Exceptional teachers	11	.8182	.405	.122

Mean Difference = -.3094

All six items were used as predictors in a discriminant functional analysis of exceptional versus "other teachers". The summary "hit table" (see Table 20) for this analysis showed that 74% of teachers were correctly classified based on these six items. The discriminant function maximizes the differences among nominal groups and may capitalize on sample-specific information. Cross-validation of these findings with a different sample of teachers might produce lower classification accuracy.

BEST COPY AVAILABLE

Table 20. **Discriminant function results for 23 specific reading strategy items including the use of systematic motivational techniques.**

Classification results -

Actual Group	No. of Cases	Predicted Group Membership 0	1
Group 0 Other teachers	56	46 82.1%	10 17.9%
Group 1 Exceptional teachers	11	2 18.2%	9 81.8%

Percent of "grouped" cases correctly classified: 82.09%

Dimensional and categorical analyses of the relationship between teacher effects and reading instruction philosophy and practices lead to a rejection of the hypothesis of independence. The variables which were dependably correlated with teacher effects and dependably distinguished "exceptional teachers" from "other teachers" included the following:

- more small group reading instruction,
- more disagreement with the notion that reading and writing develop naturally,
- more guidance of student during initial practice,
- more use of some form of published test preparation materials; and
- more use of systematic motivational strategies.

Strategies which were correlated with teacher effects in one of the analyses, but not the other, included the following:

- more teacher-directed instruction than student choice,
- more development of word attack strategies,
- more explicit and direct phonics instruction,
- more use of individual student oral reading, and
- less use of a whole language approach

BEST COPY AVAILABLE

Two null hypotheses failed to be rejected. There was no evidence to refute the hypothesis that teacher value-added is independent of teacher experience and no evidence to reject the hypothesis that teacher value-added is independent of teacher academic credits earned.

DISCUSSION

This study examined the stability of teacher effectiveness using a value-added indicator of the contributions of teachers to the reading achievement of second grade students. The multiple regression formula used to isolate teacher effects, controlled for student reading pre-test scores, gender, poverty, race, English proficiency, special education status, family composition and neighborhood poverty. In preliminary tests of the model, each of the above demographic factors contributed significantly to the prediction of second grade reading proficiency. The regression model was found to be highly robust with high cross-sample validity.

Evidence from three consecutive independent samples of continuously enrolled students demonstrates that effectiveness in reading instruction as measured by student achievement was a stable characteristic of Minneapolis teachers. Stability correlations were dependably different from zero even when the classroom effects were calculated from only seven continuously enrolled students. Median stability coefficients ranging from about .4 to .6 were consistent with earlier studies on the dependability of teacher effects in reading.

Analysis of the consistency of value-added coefficients using multi-year data in this study found considerable increase in dependability with aggregation across multiple years. The generalizability coefficient, similar to Cronbach's Alpha statistic, increased from .48 for a single year to .74 for three years' and .78 for four years' data. At a minimum, any high stakes teacher accountability system should use two year's of complete value-added data. This recommendation is consistent with recommendations from Dallas and Tennessee where the value-added systems employ two and three to five years of data respectively (see Millman, 1997 for details). In this study, the generalizability coefficient increased from .48 to .65 with two waves of reading achievement & demographic data.

Investigation of the correlations among teacher instructional behaviors and value-added teacher effects was first conducted assuming the teacher effect to be a continuous equal interval variable and later treating the teacher effect coefficient as a rank order variable used to distinguish "exceptional" teachers from "other teachers". The following discussion will first focus on the instructional behaviors that were consistent findings in both types of analysis.

More use of guided practice was correlated with higher value-added for reading. Guided practice was highlighted by Good & Brophy (1986), Rosenshine and Stevens (1986), and Carnine & Silbert (1979) as a critical aspect of effective direct reading instruction. These early studies also highlighted the amount of time actively engaged in reading groups as an important variable in effective classrooms. Teachers with higher value-added in this study reported using more small group reading instruction. This finding is somewhat inconsistent with the finding of Pressley et. al. (1996) that outstanding teachers, nominated by their reading supervisors, tended to use more whole class instruction than small group instruction. Teachers with the highest value-added in second grade reading also tended to disagree with the statement, "reading and writing develop naturally, like speaking," a central tenant of the whole language philosophy.

These results suggest that the "exceptional teachers" reported strategies which were consistent with direct instruction philosophies and consistent with the findings of National Academy of Sciences study on preventing reading difficulties in young children (Snow, et. al, 1998). Exceptional teachers in this study advocated explicit and direct skills instruction and increased individual student oral reading. They also endorsed independent reading, journal writing, encouragement of prediction while reading, and other strategies which are associated with whole language instruction. These elements of a whole language approach were reported no less by "exceptional teachers" than "other teachers".

Findings on the question which asked specifically about use of a whole language approach with selected below average students were mixed. Dimensional analysis showed a negative correlation between value-added and use of a whole language approach, but

categorical analysis found no dependable difference between "exceptional teachers" and "other teachers" in the use of whole language.

Exceptional teachers were more likely to use published test preparation material than "other teachers". However, there was no difference between the two groups in the use of expensive and time consuming curricula like "scoring high on the CAT." The obtained correlation between time spent in test preparation and teacher value-added was not dependably different from zero.

Exceptional teachers were also more likely to report the use of systematic motivational strategies for selected below average students. Teachers with the highest value-added reported use of reinforcers such as stickers, points, or special activities 82% of the time while 51% of "other teachers" reported using systematic motivational strategies.

Two negative findings are consistent with earlier teacher effects studies. Neither teacher academic credits earned nor the number of teaching years correlated dependably with value-added effects in reading.

The value-added model specified for this study has certain assumptions which present caveats to the interpretation of the teacher effects. First the model assumes a linear growth model with no interaction between teacher effects and demographic characteristics. It also assumes no interaction among teacher effects. Students instructed in reading by more than one teacher (e.g. a special education resource teacher in addition to the classroom teacher) do not have estimates for both teachers in this model. Teachers involved in team teaching of reading were excluded from the analysis.

In addition to the above considerations are a number of limitations associated with the particular measurement instruments and procedures use in this study. In particular, the standardized reading comprehension and vocabulary tests scores available through the districtwide assessments may not reflect all relevant aspects of second grade reading. The lack of constructed response items may restrict the measurement of reading comprehension. The omission of word analysis subtests and direct measures of fluency may also limit the validity of teacher effects in reading. The district decision to eliminate the vocabulary subtest in Year 3 (1995-96) not only limited the generalizability of the findings but also negatively affected the reliability of the reading post-test measure.

The value-added coefficient estimated in this study is limited by the available student and family characteristic variables coded in the district central computer system. This coefficient may be biased due to missing student demographics, school characteristics or neighborhood variables. In particular, free or reduced price lunch status and residential zip codes may be weak proxies for family income and education. Numerous studies have documented the high correlation between achievement levels and median family income and mother's education. The lack of these variables may bias the value-added coefficient and identify teacher effects that are at least partially confounded with family involvement, achievement expectations and addition assistance available to middle class parents, but less prevalent with families in poverty.

Conclusions And Recommendations

The results of this study corroborate findings from previous generations of research that teacher effects in early reading are relatively stable. Stability coefficients for two year data ranging from .4 to .6 are somewhat higher than coefficients found in the Brophy studies in the 1970s. This could be due in part to the use of a full complement of individual student regressors in the prediction model which isolated the teacher effects from individual student and family characteristics. It may also be due in part to greater variability in reading instruction in the 1990s.

In order for teacher effects to be accepted as unbiased and accurate indicators of reading instruction efficacy, teacher effect calculations should include control for prior learning and correlated factors not under the influence of the teacher. Value-added indicators, such as the one used in this study, may provide a more defensible method for distinguishing "exceptional teachers" from "other teachers" than the use of student gain measures alone.

This study provides support for skills-based instruction in early reading. Teachers who "beat the odds" in this study tended to endorse more direct instruction activities, including greater use of teacher guidance during initial instruction and more use of small group instruction. Teachers identified as "exceptional" through value-added analysis endorsed more teacher directed activities, more development of word attack strategies, more explicit and direct phonics instruction, and more use of individual student oral reading.

This study also found dependable relationships between the use of test preparation activities and teacher effects. However, use of expensive class-period-long test preparation curricula had no measurable advantage. Teachers who reported using systematic motivational techniques with below average students had higher overall value-added effects for reading.

The question of missing variable bias also needs to be addressed in future research. Does the lack of a strong socio-economic indicator (e.g. family income) fail to adequately represent the contributions of the family to student learning? Would the inclusion of a mother's education variable significantly change the teacher value-added estimate? Use of more sophisticated statistical models could also differentiate value-added effects for certain types of students. Do some teachers provide higher value-added for students who are below average while other teachers provide higher value-added for students who are above average? What are the characteristics of teachers who produce high value-added for both groups of students? These questions, raised in an article by Reynolds & Heistad (1997) could be more fully developed and investigated using value-added statistical procedures.

It would be interesting to replicate this study of second grade reading teacher effect stability with an oral reading pre-test and post-test. Preliminary evidence suggests that an oral reading performance measure had equal predictive validity to a standardized paper and pencil test of reading comprehension in first and second grades (Heistad, 1998). Would the same teachers be identified as exceptional using different dependent variables? Should multiple dependent variables be considered for teacher accountability systems?

Future research on value-added correlates should also include more in-depth measures of instructional behaviors taken from interview and direct classroom observation similar to the on-going studies of Pressley et al. (1996). Perhaps future studies would better serve teachers and researchers if they focused not on Whole Language versus Direct Instruction approaches but on how exceptional teachers implement balanced instruction curricula and methods in their classrooms. Classroom observation methodology could also focus on student motivation and classroom management issues which have been important issues for decades (Freiberg et al., 1995) and re-surfaced in this study.

Based on the challenges in implementing a high stakes teacher accountability system which does not have unintended side-effects, this investigator recommends against using teacher value-added analysis to pay teachers directly for higher reading test scores.[5] Recognizing and holding in high esteem those teachers who "beat the odds" should be considered instead. The teachers of students who excel, despite personal histories and demographics which would predict otherwise, should be considered human capital. These teachers should be highly valued as mentors, models for emulation, and subjects for in-depth investigation. They should be given the opportunity to tell their story to colleagues and the general public. This type of reward system would, I believe, contribute to the professionalization of teaching. It also has a built-in validity check. Teachers who are given distinction must open their classroom doors to observers, demonstrate their wares and not simply cash a bonus check.

---

[5] See Kelly, 1997; Hannushek & Jorgenson, 1996; Odden & Kelly, 1997; & Walberg & Paik, 1997 for different perspectives on this subject.

Another possibility is to provide an incentive for exceptional teachers to accept assignments to schools with the most at-risk students. Current seniority and teacher bidding guidelines result in a concentration of less skilled teachers in the highest poverty schools in Minneapolis. A corps of exceptional teachers might be used to "turn the reading program around" in schools where lack of student progress may put the school at-risk for reconstitution as provided in the local teachers contract (Minnesota Federation of Teachers, 1997).

Accountability for early reading mastery should not be focused solely at the school or district level. It is at the classroom level with individual teachers where the cost-benefit analysis will have its greatest effect. From Benjamin Bloom's (1964) classical analysis of stability and change in human behavior comes this quote on the critical need for effective early reading instruction,

> "...approximately 50% of general achievement at grade 12 has been reached by the end of grade 3 (age 9). This suggests the great importance of the first few years of school was well as the preschool period in the developing of learning patterns and general achievement. These are the years in which general learning patterns develop most rapidly, and failure to develop appropriate achievement and learning in these years is likely to lead to continued failure or near failure throughout the remainder of the individual's school career. The implications for more powerful and effective school environments in the primary school grades are obvious." (p.127)

The past 35 years of teacher effects research confirms what Bloom recognized. Teachers can and do have differentially large positive or negative effects on the early reading success of their students. This study has shown that these environmental effects are stable across time for individual teachers and that those teachers who are most successful in early reading instruction can be identified empirically.

BEST COPY AVAILABLE

REFERENCES

Acland, H. (1976). Stability of teacher effectiveness: A replication. Journal of Educational Research, 69, 289-292.

Adams, M.J. (1990). Beginning to Read. Cambridge MA: Harvard University Press.

Adams, M.J. (1996). Adventures in mind-tuning: From thinking skills to enlightened phonics. Sylvia Scribner Address to the Annual Meeting of the American Educational Research Association.

Bembry, K, Jordon, H.R., Gomez, E., Anderson, M.C., & Mendro, R.L. (1998). Policy implications of long-term teacher effects on student achievement. Paper presented at the annual meeting of the American Education Research Association, San Diego, April.

Bembry, K., Dash, W., & Mendro, R. (1997). Classroom effectiveness indices: Statistical methodology, post-hoc analysis, and practical applications. Paper presented at the annual meeting of the American Education Research Association, Chicago, March 24-28.

Berk, R.A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. Journal of Personnel Evaluation in Education, 1, 345-363.

Berliner, D.C., & Biddle, B.J. (1995). The manufactured crisis: Myths, fraud, and the attack on America's public schools, New York: Longman.

Bloom, B. (1964). Stability and change in human characteristics. New York, Wiley.

Bock, R.D. & Wolfe, R. (1996). Audit and review of the Tennessee value-added assessment system (TVAAS): Final report. A Review and Analysis of the Tennessee Value-Added Assessment System (Part 1). State of Tennessee Comptroller of the Treasury, Office of Educational Accountability (obtainable via telephone: 615-532-1111).

Brennan, R.L. (1983). Elements of generalizability theory. Iowa City: American College Testing Program.

Brophy, J. & Good, T. (1986). Teacher behavior and student achievement. In M.C. Wittrock (Ed.), Handbook of research on teaching (3rd ed., pp. 328-325). New York: Macmillan.

BEST COPY AVAILABLE

Brophy, J. (1973). Stability of teacher effectiveness. American Education Research Journal, 10, 245-252.

Brophy, J. (1988). Research on teacher effects: Uses and abuses. Elementary School Journal, 89, 3-22.

Brown, J.L. (1971). Stability of teacher effectiveness in the teaching of reading at the first grade level. Unpublished doctoral dissertation, University of Minnesota.

Bryk, A.S. & Raudenbush, S.W. (1992). Hierarchical linear models. Newbury Park, CA: Sage.

Carnine, D., & Silbert, J. (1979). Direct instruction reading. Columbus, OH: Merrill.

Chall, J.S. (1983). Stages of reading development. New York: McGraw Hill.

Chall, J.S., Jacobs, V.A., & Baldwin, L.E. (1990). The reading crisis: Why poor children fall behind. Cambridge, MA: Harvard University Press.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach, L.J. & Furby, L. (1970). How we should measure change - - or should we? Psychological Bulletin, 74, 68-80.

Emmer, E., Evertson, C., & Brophy, J. (1979). Stability of teacher effects in junior high classrooms. American Education Research Journal, 16, 71-75.

Foorman, B.R. (1995). Research on "the great debate": Code-oriented versus whole language approaches to reading instruction, School Psychology Review, 24, 1995.

Foorman, B.R., Francis, D.J., Fletcher, C., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. Journal of Educational Psychology, 90, 37-55.

Freiberg, H.J., Stein, T.A., & Huang, S. (1995). Effects of a classroom management intervention on student achievement in inner-city elementary schools. Educational Research and Evaluation, 1, 36-66.

Furhrman, S., & O'Day, J. (1996). Rewards and reform: Creating educational incentives that work. San Francisco, Jossey Bass.

BEST COPY AVAILABLE

Glass, G.V. (1990). Using student test scores to evaluate teachers. In J. Millman & L. Darling-Hammond (Eds.), The new handbook of teacher evaluation (pp. 191-215). Newbury Park, CA: Sage.

Good, T. (1979). Teacher effectiveness in elementary school: What we know about it now. Journal of Teacher Education, 30, 52-64.

Goodman, K. (1986). What's whole in whole language? Portsmouth, NH: Heineman

Goodman, K.S. (1992). Whole language research: Foundations and development. In S.J. Samuels & A.E. Farstrup (Eds.), What research has to say about reading instruction, vol. 2, 46-69, Newark Delaware: International Reading Association.

Goodman, K.S. and Goodman, Y.M. (1979). Learning to reading is natural. In L.B. Resnick and P.A. Weaver (Eds), Theory and Practice of Early Reading, vol. 1, 137-154. Hillsdale, NJ: Erlbaum Associates.

Gough, P.B. & Hillinger, M.L. (1980). Learning to read: An unnatural act. Bulletin of the Orton Society, 30, 179-196.

Hanushek, E.A., & Jorgenson, D.W.. (1996). Improving America's Schools: The role of incentives. National Research Council. Washington, DC: National Academy Press.

Heistad, D.J. (1998). A district-wide individually administered oral reading assessment in first grade: Technical adequacy and standards-based reporting. Paper presented to American Education Research Association annual conference, San Diego, April 14.

Koretz, D. (1996). Using student assessments for educational accountability, in E.A. Hanushek & D.W. Jorgenson (Eds.), Improving America's Schools: The Role of Incentives, 171-196, Washington D.C.: National Academy Press.

Mandeville, G. K., & Anderson, L.W. (1987). The stability of school effectiveness indices across grade levels and subject areas. Journal of Educational Measurement, 24, 203-216.

Mandeville, G.K. (1988). School effectiveness indices revisited: Cross-year stability. Journal of Educational Measurement, 25, 349-356.

Mandeville, G.K., & Rivers, J. (1991). The South Carolina PET study: Teachers' perceptions and student achievement. Elementary School Journal, 91, 377-407.

BEST COPY AVAILABLE

Meyer, R.H. (1994). Can schools be held accountable for good performance? A critique of common educational performance indicators. Working paper, Harris Graduate School of Public Policy Studies, University of Chicago.

Meyer, R.H. (1996). Value-added indicators of school performance, in E.A. Hanushek & D.W. Jorgenson (Eds.), Improving America's Schools: The Role of Incentives, 171-196, Washington D.C.: National Academy Press.

Millman, J. (1997). Grading teachers, grading schools: Is student achievement a valid evaluation measure. Thousand Oaks, CA: Corwin.

Minnesota Federation of Teachers (July, 1997). Teacher Contract, Agreements & Policies. Unpublished labor agreement. Minneapolis Public Schools: author.

Odden, A., & Kelley, C. (1997). Paying teachers for what they know and do: New and smarter compensation strategies to improve schools, Thousand Oaks, CA: Corwin.

Olson, L. (1998). A question of value. Education Week, May 13, 1998.

Pressley, M. & Rankin, J. (1994). More about whole language methods of reading instruction for students at risk for early reading failure. Learning Disabilities Research & Practice, 9, 156-167.

Pressley, M. (1994). Commentary on the ERIC whole language debate. In C.B. Smith (Moderator), Whole language: The debate (pp. 155-178). Bloomington, IN: ERIC/REC.

Pressley, M., Rankin, J., & Yokoi L. (1996). A survey of instructional practices of primary teachers nominated as effective in promoting literacy. Elementary School Journal, 96, 363-384.

Raudenbush, S.W., & Willms, D.J. (1995). The estimation of school effects, Journal of Educational and Behavioral Statistics, 20, 307-336.

Reynolds, M.C. & Heistad D.J. (1997). 20/20 Analysis: Estimating school effectiveness in serving students at the margins. Exceptional Children, 63 (4), 439-449.

Rosenshine, B, & Stevens, R. (1986). Teacher behavior and student achievement. In M.C. Wittrock (Ed.), Handbook of research on teaching (3rd ed., pp. 376-391). New York: Macmillan.

BEST COPY AVAILABLE

Sanders, W.L., & Horn, S.P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment, Journal of Personnel Evaluation in Education, 8, 299-311.

Sanders, W.L., & Horn, S.P. (1995). Educational Assessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes, Education Policy Analysis Archives, 3, 1-19. [on the web at LISTSERV@asuvm.inre.asu.edu].

Sanders, W.L., Saxton, A.M., & Horn, S.P. (1997). The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.) Grading Teachers, Grading Schools, Thousand Oaks, CA: Corwin.

Shaywitz, S.E., Escobar, B.A., Shaywitz, B.A., Fletcher, J.M. & Makuch, R. (1992). Evidence that dyslexia may represent the lower tail of a normal distribution of reading ability. New England Journal of Medicine, 326, 145-150.

Slavin, R.E. & Fashola, O.S. (1998). Show me the evidence: Proven and promising programs for America's schools. Thousand Oaks, CA: Corwin.

Snow, C.E., Burns, S., & Griffen, P. (1998). Preventing reading difficulties in young children. National Research Council. Washington, DC: National Academy Press.

Stahl, S.A., & Miller, P.D. (1989). Whole language and language experience approaches for beginning reading: A quantitative research synthesis. Review of Educational Research, 59, 87-116.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. Reading Research Quarterly, 21, 360-407.

Thum, Y. & Bryk, A. (1997). Value-added productivity indicators: The Dallas system. In J. Millman (Ed.) Grading Teachers, Grading Schools, Newbury Park, CA: Corwin.

Webster, W.J. & Mendro, R.L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.) Grading Teachers, Grading Schools, Thousand Oaks, CA: Corwin.

Webster, W.J. & Olson, G.H. (1988). A quantitative procedure for the identification of effective schools. Journal of Experimental Education, 56, 213-219.

Wittrock, M. (1986). Handbook of research on teaching (3rd ed.). New York: Macmillan.

BEST COPY AVAILABLE